Determining the Presence of Metabolic Pathways using Machine Learning Approach

Yara Saud Aljarbou[1]
College of Computer Sciences and Engineering
Taibah University, Madinah
Saudi Arabia

Fazilah Haron[2]
College of Computer and Cyber Sciences
Prince Muqrin University, Madinah
Saudi Arabia

***Abstract*—The reconstruction of the metabolic network of an organism based on its genome sequence is a key challenge in systems biology. One of the strategies that can be used to address this problem is the prediction of the presence or the absence of a metabolic pathway from a reference database of known pathways. Although, such models have been constructed manually, obviously such a method cannot be used to cover thousands of genomes that has been sequenced. Therefore, more advanced techniques are needed for computational representation of metabolic networks. In this research, we have explored machine learning approach to determine the presence or the absent of metabolic pathway based on its annotated genome. We have built our own dataset of 4978 instances of pathways. The dataset consists of 1585 pathways with each having 20 different representations from 20 organisms. The pathways were obtained from the BioCyc Database Collection. The pathway dataset also consists of 20 features used to describe each pathway. In order to identify the suitable classifier, we have experimented five machine learning algorithms with and without applying feature selection methods, namely Decision Tree, Naive Bayes, Support Vector Machine, K-Nearest Neighbor and Logistic Regression. Our experiments have shown that Support Vector Machine is the best classifier with an accuracy of 96.9%, while the maximum accuracy reached by the previous work is 91.2%. Hence, adding more data to the pathway dataset can improve the performance of the machine learning classifiers.**

***Keywords*—Metabolic pathway prediction; pathway dataset; metabolic network of organism; machine learning; support vector machine**

I. Introduction

Constructing a comprehensive model of metabolic reaction networks which occurs within every organism is a key step toward understanding the metabolism of an organism. The availability of metabolic networks as predictive tools is fundamental in many research fields such as metabolic engineering, diagnostic medicine, pharmacology, biochemistry, biology and physiology. An exciting example of using metabolic networks is the screening of disease-specific biomarkers that can be applied for early detection of diseases.

Although several of metabolic network models have been constructed through manual processing, such an approach obviously cannot be used when we have thousands of sequenced genomes. Therefore, more advanced techniques are needed for computational representation of metabolic networks.

The pathway prediction is one of the methods that can help in the reconstruction of an organism's metabolic network from its genome sequence. In the pathway prediction problem: by having the annotated genome of an organism and its reactome, we can predict the set of metabolic pathways present in the organism. In this research, by taking the reactome as predetermined by other methods, we can focus on developing an improved pathway prediction methods.

The pathway prediction can involve predicting novel pathways that have not been previously observed which also called pathway discovery, or predicting pathways that were previously known in other organisms. Our methodology does the latter, predicting pathways from a selected reference database.

Our main motivation is to develop a more accurate method for predicting metabolic pathways and to overcome the limitations of the current existing methods. It is expected that machine learning will help provide effective knowledge from a variety of big data including data about metabolism.

The aim of our research is to improve the accuracy of the metabolic pathway prediction process. The outcome of this research is a comparative study of our chosen machine learning algorithms with the work that have been established in this domain.

The rest of the paper is organized as follow: Section II provides a background knowledge of our research, and presents the existing researches that are related to our research. Section III provides the details about the methodology that has been adopted in this research. Section IV describes the experimental results obtained from our experiment. Section V summarizes the whole research and provides some suggestions on how the work could be further improved.

II. Background and Literature Review

This section provides a background knowledge of our research, and also presents the works that are related to our research.

A. Backgound of Metabolic Pathways

Bioinformatics is a rapid development research area for the scientific community. The bioinformatics research focuses on algorithms, statistical approach, computations approach and developing huge databases to solve problems in the biology

field. One of major research efforts in this field is metabolic pathway prediction.

In biochemistry, the metabolism is a series of chemical reactions that take place within the cells of organisms based on enzymes; which are necessary to ensure that the organism survives [1]. Enzymes in the process play an active role in the reproduction and growth of living organisms and they stimulate the organisms to interact with and respond to their environment. It also said that the term metabolism, referred to all the biochemical processes carried out by the bodies of the biological organisms starting from the production of new tissues based on basic nutrients breaking down of carbohydrates, sugars and fats then turning them into energy for the body to carry out daily activities. There are two main goals for metabolism [2]: the first is gaining energy that enables the cell to perform its functions through demolition reactions, which also known as Catabolism, and the other goal is compounding complex organic compounds that are necessary for the cell through building reactions, which also known as Anabolism.

These two goals rarely achieved by a single chemical reaction. Rather, the dominant rule is to produce energy or synthesize the compounds through a number of successive reactions so that the material from the first reaction is a reactive substance in the second reaction. A set of reactions that transform a specific material into another material called metabolic pathway. A metabolic pathway has many steps, each step begins with a specific molecule and ends with a product, and each reaction is catalyzed by a specific enzyme as shown on the Fig. 1.

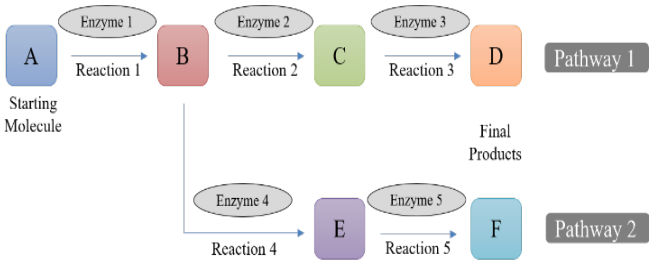

Fig. 1. Metabolic Pathways.

There are three main goals for studies conducted on metabolic pathways [3]: (1) Identification of intermediate compounds and enzymes involved in a series of reactions that positively understand the activation and inhibition of this path. (2) Identify the mechanism of an interaction in a series of reactions that requires the separation of enzymes that help in each reaction. (3) Identify the mechanism of regulating the speed of different reactions.

The principal axiom of Systems Biology is that a system should be also analyzed at the level of interactions of its parts, not only as sum of them.

B. Related Work

There have been many previous works on solving the metabolic pathway prediction problem, such as the computational method that was developed to compare organisms based on genome-wide metabolic pathway analysis [4], using the WIT (What Is There) database, which is a metabolic pathway profile for each completed genome [5]. These profiles are records of the presence and absence of the various metabolic pathways, and constitute the basis for a comparison of organisms. The developed methodology requires that all the reactions in a pathway to have enzymes in order for the pathway to be consider present.

KEGG project on “pathway maps” based on the information of the genome [6]. KEGG pathway maps encompassed varied metabolic pathways from varied organisms. One of the issues faced in the project is the problem of pathway map prediction rather than the problem of pathway prediction. The description of KEGG’s algorithm for map prediction and the accuracy evaluation of that algorithm are no ware to be found.

Matthews1 et al. [7], performed prediction of metabolic pathways based on the information of the genome stored in the Reactome Knowledgebase, which is an online, manually curated resource that provides an integrated view of the molecular details of human biological processes that range from metabolism to DNA replication and repair to signaling cascades [8]. However, the description of their algorithm and the accuracy evaluation of their algorithm are no ware to be found.

In [9] and [10], Kastenmüller et al. developed an outcome similar to the “information content” features used in the predictors of [11], calculating the fraction of reactions present in the pathway, weighted in terms of the unity of the reaction. It is expected that such analyses could be improved by taking advantage of the probabilities of pathway presence.

Al Daoud [12] developed a new algorithm to predict pathway classes and individual pathways for a previously unknown query molecule. His main idea was to use a dense graph, where the enzymes are represented as edges and the compounds as vertices. The weights are assigned to the edges according to the previous known pathways. He applied the shortest path algorithm for each missing enzyme in a pathway. A pathway is considered to be belong to an organism if the total cost between the initial and final compound is higher than a threshold. The validation of their experiments showed that the suggested algorithm is capable to classify more than 90% of the pathways correctly.

None of the above work involve in the predicting the absence or presence of pathway. There is PathoLogic [13], which is a known tool that can be used to predict the presence or the absence of metabolic pathways in sequenced and annotated genomes.

Another highly related work to ours is Dale et al. [11], they developed a machine learning methods for metabolic pathway prediction, their method showed a better performance when compared with the standard methods presented in the hard-coded pathway prediction tool PathoLogic [13], while at the same time allowing easier explanation, tenability, and extensibility of the results. Table I summarizes a comparison between the PathoLogic tool and the machine learning approach.

TABLE I. COMPARISON BETWEEN PATHOLOGIC AND ML APPROACH

PathoLogic Algorithm [13]	ML Algorithm (Logistic Regression) [11]
Accuracy of 91%	Accuracy of 91.2%
No additional information	Provide probability for each predicted pathway
Requires Experts	No Experts Required
Developed and refined over approximately a decade	Developed with well-designed collection of input features

Machine learning algorithms that has been tested in pathway prediction included logistic regression, decision trees and naive bayes. The goal of this study was to test different machine learning methods for the determination of the presence or the absence of a metabolic pathway based on the pathways information for many organisms presented in the pathway collection MetaCyc and to develop new predictors for determining the presence of a metabolic pathway in newly sequenced organisms. In order to evaluate their methods, Dale et al. have developed the gold standard pathway dataset [11].

III. METHODOLOGY

This section provides the details about the methodology that has been adopted in this research.

The overall research methodology was based on the machine learning process. The methodology is divided into four phases, namely data acquisition, data preprocessing, training, and model evaluation. We decided to use Weka, which is one of the most popular and freely available machine learning tool [14].

The first phase of our methodology was the data acquisition phase, in which we collected the relevant data (pathways information) from the BioCyc databases for the study and computed the values of the features based on their description. The data preprocessing phase was the second phase, in which the collected data was cleaned and integrated such that, the datasets were proper for the process of classification. We also applied feature selection methods on the dataset to select the most effective feature to train our machine learning algorithms. The data from the second phase which is the data preprocessing phase were then passed over to the third phase. The third phase is the training of the machine learning algorithms, which consists of two parts; without feature selection and with the selected feature from second phase. In the fourth phase, the model evaluation and comparison phase, we have tested the classifiers without feature selection and the classifiers that used the selected features by using standard 10-fold cross validation. We also performed a comparative analysis between the different classifiers based on widely used evaluation metrics. Fig. 2 represent the overall methodology of our research. Each phase of our research methodology will be explained in detail in the following sections.

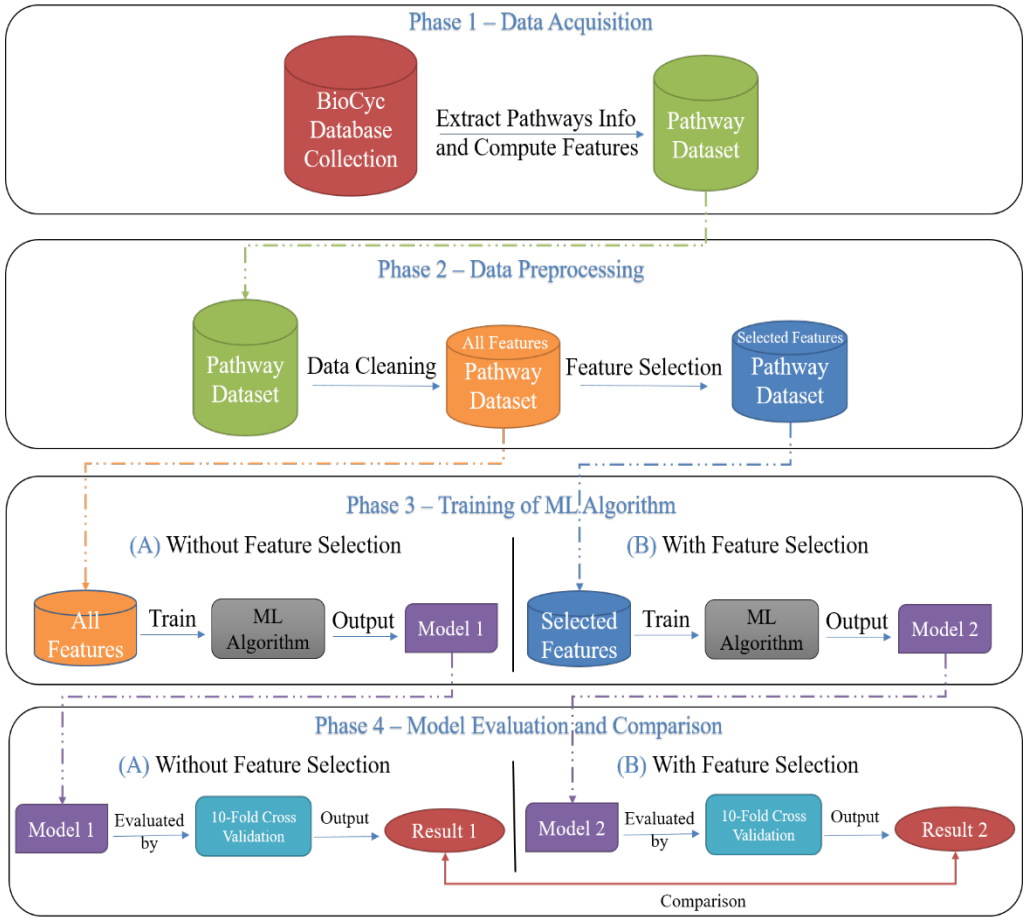

Fig. 2. Overall Research Methodology.

A. Data Acquisition Phase

In order to train our machine learning methods and validate them against each other, we have constructed a pathways dataset containing known information about the presence or absence of pathways in different organisms. Here we describe the construction and content of the dataset.

First, we accessed the pathways databases from the BioCyc Database Collection, which is a collection of 14560 Pathway/Genome Databases (PGDBs) [15]. We then obtained the list of features and their description that has been used by Dale et al. in [11]. The pathways dataset currently contains 4979 elements that describe pathway absence and presence in 20 organisms as shown in Table AI in the Appendix, along with the corresponding Pathway/Genome Databases (PGDBs) [15]. All our data were mainly derived from PGDBs. Each element or instance of the dataset is a triple of the form (Organism Name, Pathway ID, Is-Present?), as shown in Table II.

TABLE II. ATTRIBUTE DESCRIPTION OF PATHWAY DATASET

Attribute	Description
Organism Name	The organism's name.
Pathway ID	The pathway's ID.
Is-Present?	The presence or absent of the pathway (Label).

Organism's name, is not considered as a factor that determines the presence or the absent of a particular pathway. We only use it to determine the number of instances (different representation) of a pathway. For example, we have 20 different representations of pathway 1 from 20 different organisms.

We used the MetaCyc Metabolic Pathway Database, which is "a curated database of experimentally elucidated metabolic pathways from all domains of life, as a reference for the curated metabolic pathways, MetaCyc contains 2801 pathways from 3123 different organisms" [16]. In order determine the presence or the absent of the pathways in each organism, we applied the two rules. For each organism, the first rule is to mark as positives (present) all the pathways that are present in the database corresponding to each organism [11]. Then, we added as negatives (absent) all pathways in the same databases but has not been annotated in MetaCyc database [11]. As for the second rule, we added as negatives (absent) all the pathways that have no enzymes [4]. As for features that describe each pathway, we used some of the features that have been defined and used by Dale et al. [11] due to the absent of expert in this domain. Table AII in the Appendix describes the features that we extracted and used in our research.

At the end of this phase, we have a complete version of the pathway dataset. The pathway dataset has 4979 instances and 22 attributes including the ID of the pathways, the label of each pathway (present or absent) and finally the 20 features that describe each pathway. The number of unique pathways in the pathway dataset is 1585 pathways with each having 20 different representations from 20 organisms as compared to the original dataset that has been constructed by Dale et al. [11], which only have 6 organisms.

After describing each part of the pathway dataset, Table III shows a sample of the complete pathway dataset. The "…" in Table III refers to the remaining features. In the sample, we only mentioned two of the 20 features, the Biosynthesis-Pathway and the Num-Reactions features.

B. Data Preprocessing Phase

The data preprocessing stage, were the collected data from the data acquisition phase was integrated, and then the data went through the cleaning process, in which we deleted unfitting entries of the data, such as those that provide unrelated results in the dataset.

We also applied feature selection methods on the pathway dataset in order to get the most effective features in determining the presence or the absent of each pathway in the pathway dataset. Feature selection methods can reduce both the computational complexity and the data in the dataset. Therefore, the dataset can also be more useful and efficient to train the classification algorithms [17]. One of the feature selection methods that we used is Information Gain (IG), "which measures how much 'information' a feature gives us about the class or the label. Features that perfectly partition should give maximal information and produce a higher score than the unrelated features that give us no information about the value of the class or the label" [18]. The other feature selection method that we used is Correlation-Based Feature Selection (CFS), which "evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them, the features subsets that are highly correlated with the class while having low inter-correlation are preferred" [19].

TABLE III. SAMPLE OF PATHWAY DATASET

Organism Name	Pathway ID	Features			Is-Present?
		Biosynthesis-Pathway	…	Num-Reactions	
Escherichia coli	CYANCAT-PWY	FALSE	⋯	3	PRESENT
Escherichia coli	ARO-PWY	TRUE	⋯	0	ABSENT
Arabidopsis thaliana	PWY-3781	FALSE	⋯	4	PRESENT
Arabidopsis thaliana	PWY-6754	FALSE	⋯	2	ABSENT
Bacteroides thetaiotaomicron	GLYSYN-PWY	FALSE	⋯	1	PRESENT
Bacteroides thetaiotaomicron	PWY-7353	FALSE	⋯	2	ABSENT

In our research, we have used both of these feature selection methods in order to get the most effective features in determining the presence or the absent of a particular pathway. More details about the experiment will be discussed later in later section.

C. Training of Machine Learning Algorithm Phase

In this research, five commonly used classification algorithms [20], namely, K-Nearest Neighbor, Decision Tree, Naive Bayes, Logistic Regression and Support Vector Machine were evaluated. The first four algorithms were chosen since the purpose of the work is to compare with the work in [11], where they use the same algorithms. Support Vector Machine was chosen because it is also one of the most widely used classification algorithms.

D. Model Evaluation and Comparison Phase

For evaluating the performance of the prediction techniques, we used several performance measures that are widely used. The method for evaluating the classification models was checking the confusion matrix. The confusion matrix contains information about the predicted and the actual classifications that we get from the proposed classifier [21]. The other evaluation metrics assessed for effectiveness measurement were classification accuracy, specificity, and sensitivity or recall [21].

In order to test the models that we got from the training of machine learning algorithm phase, we used the k-fold cross validation. In this test, the dataset is randomly divided into K equal parts, one part is selected to test the model, and the k-1 is the remaining part used to train the model [22]. In our research, we used the standard 10-folds cross validation to test and evaluate our models. The main advantage of this method is that it is simple to implement and does not require much time in the calculation process.

IV. Experimental Results and Discussion

In this section we describe the experimental results that we got from the last three phases of the research methodology, and these are data preprocessing, training of machine learning algorithm, and model evaluation and evaluation.

A. Feature Evaluation and Selection

In the data preprocessing phase, we have applied two feature selection methods in order to reduce both the computational complexity and the data in the pathway dataset.

The first method is information gain, which measures how much “information” a feature gives us about the class or the label, and base on this information, the information gain method assigns scores to the features and rank them based on these scores. Based on these scores, we started to remove the features with the lower score one by one and at each time, we used the remaining features to train the machine learning algorithm and observed its accuracy. The machine learning algorithm that we used in this experiment is the Naive Bayes. The results of this experiment are shown in Fig. 3.

The results showed that by using the feature with the higher score, we reached the highest accuracy which is 92.7%, while when using the two features with highest score, the accuracy reduced to 89.4%. From this, we can say that the gain information method does not tell us what are the best combination of features to reach higher accuracy, therefore; another feature selection method is needed.

The other feature selection method that we have used in our experiment is the correlation-based feature selection, which evaluates the worth of a subset of features by considering the individual predictive ability of each feature along with the degree of redundancy between them. The best subset of features that has been selected by this method is a combination of the features Deg-Or-Detox-Pathway, Is-Sub-Pathway, Has-Enzymes and Has-Key-Reactions. By using these features to train our machine learning algorithms, we reached a higher accuracy level.

B. Evaluation of Classification Models without Feature Selection

In the first part of the training of machine learning algorithm phase, we constructed five classification models using the pathway dataset for the Naive Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbor and Support Vector Machine classifiers. The models were constructed by using all the features in the pathway dataset without applying any feature selection methods. We used the standard 10-fold cross-validation for testing our models in the model evaluation and comparison phase.

Fig. 4 shows the results of the experiment, the accuracy, sensitivity, and specificity rate of the classification models for the Naive Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbor and Support Vector Machine classifiers.

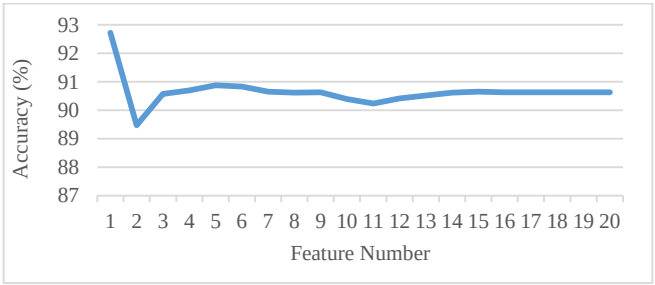

Fig. 3. Accuracy (%) of Naive Bayes with Feature Selection by Information Gain Method.

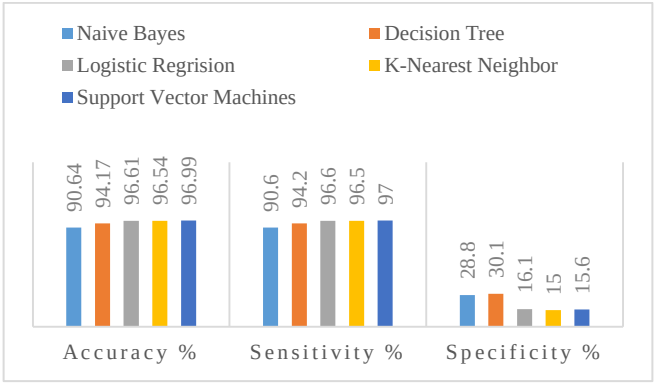

Fig. 4. Classification Models’ Performance without Feature Selection.

The experiment results indicated that the accuracy percentage of the pathway classification model that used Support Vector Machine as the classifier gave 96.9%, which is the highest compared to the Naive Bayes, Decision Tree, Logistic Regression and K-Nearest Neighbor. In this experiment, we did not apply any feature selection methods.

The sensitivity, which is the proportion of the present pathways classified as present, we can see that the classifier of the Support Vector Machine gave the highest sensitivity of 97%. As for the specificity, which is the proportion of the absent pathways classified as absent, the classifier that gave the highest specificity of 30.1% was the Decision Tree. Therefore, we can say that among the five classifiers that we have trained and evaluated, the best classifier in predicting the present pathways is the Support Vector Machine and the Decision Tree is the best in predicting the absent pathways.

C. Evaluation of Classification with Feature Selection

In the second part of the training of machine learning algorithm phase, we constructed the same five classification models using the pathway dataset for the Naive Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbor and Support Vector Machine classifiers, but this time, we constructed the models by using the selected features by the feature selection methods. We used the standard 10-fold cross-validation for testing our models in the model evaluation and comparison phase.

Fig. 5 shows the results of the experiment, the accuracy, sensitivity, and specificity rate of the classification models for the Naive Bayes, Decision Tree, Logistic Regression, K-Nearest Neighbor and Support Vector Machine classifiers, which are trained by the pathway dataset that only contains the selected features.

The experiment results indicated that the accuracy percentage of the pathway classification model that used Support Vector Machine as classifier gave 96.9%, which is the highest compared to the Naive Bayes, Decision Tree, Logistic Regression and K-Nearest Neighbor. The pathway dataset used in this experiment only contained the features selected by the feature selection methods.

The sensitivity, which is the proportion of the present pathways classified as present, we can see that the classifier of the Support Vector Machine gave the highest sensitivity of 97%. As for the specificity, which is the proportion of the absent pathways classified as absent, the classifier that gave the highest specificity of 30.1% was the Decision Tree. Therefore, we can say that among the five classifiers that we have trained end evaluated, the best classifier in predicting the present pathways is the Support Vector Machine and the Decision Tree is the best in predicting the absent pathways.

D. Classification Models Comparison with and without Feature Selection

In the model evaluation and comparison phase, a comparative analysis of the constructed models with and without feature selection was performed.

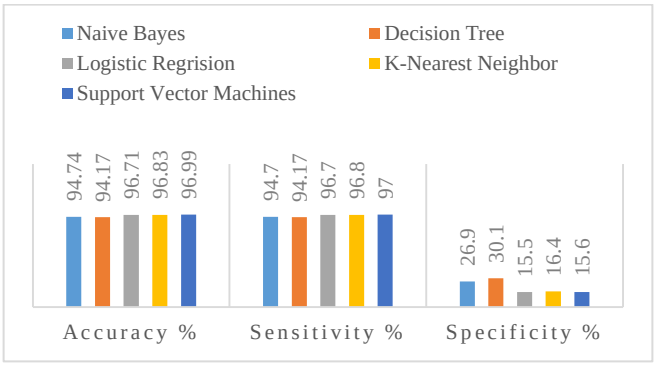

Fig. 5. Classification Models' Performance with Feature Selection.

The results show that by applying a proper feature selection method, some of the classifiers gets higher results than their corresponding classifiers without feature selection in terms of the accuracy level and while others got the same accuracy level. However, none of the classifiers produces results that are lower. The best result for pathway prediction in terms of accuracy, was given by the Support Vector Machine classifier, which obtained 96.9% accuracy, followed by K-Nearest Neighbor with 96.826 accuracy, Logistic Regression with 96.7% accuracy, Naive Bayes reaching the 94.7% accuracy, and Decision Tree with 94.1% accuracy level.

However, in the experiments without using any feature selection methods, The best result for pathway prediction in terms of accuracy, was given by the Support Vector Machine classifier which obtained 96.9% accuracy, followed by Logistic Regression with 96.6% accuracy, K-Nearest Neighbor with 96.5% accuracy, Decision Tree with 94.1% accuracy, and Naive Bayes reaching the 90.6% accuracy level. Fig. 6 shows a graphical representations of the results.

E. Comparative Analysis of our Models and the Existing Models

In this section, a comparative analysis between the models constructed by the pathway dataset that we build and the models constructed by the pathway dataset in [11] was performed. The machine learning classifiers used by Dale et al. [11] are Naive Bayes, Decision Tree, Logistic Regression and K -Nearest Neighbor, but they did not include the results for the K-Nearest Neighbor classifier. Therefore, we only compared the accuracy of the first three classifiers with the accuracy given by the classifiers constructed by our pathway dataset. A graphical representation of the comparison results based on the accuracy level of our classification models and the existing classification models shown in Fig. 7.

From Fig. 7, we can see that the models constructed by our pathway dataset out-performed the models in [11]. The Logistic Regression classifier in both experiments gave us the highest accuracy compared to the Decision Tree and the Naive Bayes classifiers.

In our research, we also constructed Support Vector Machine classifier, which gave us a higher accuracy equal to 96.9%, while the K-Nearest Neighbor classifier gave slightly lower, that is 96.8%.

Fig. 6. Classification Models' Accuracy (%) with and without Feature Selection.

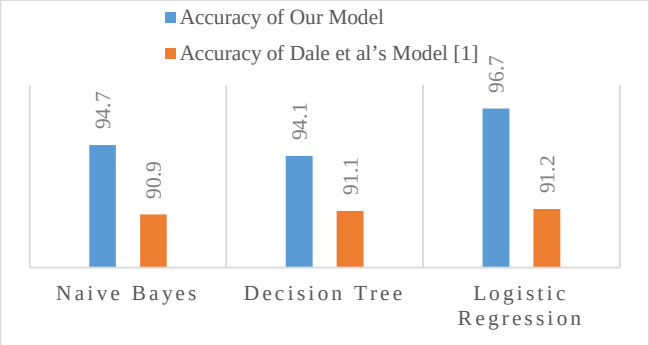

Fig. 7. Accuracy (%) of our Models and the Existing Models.

It is to be noted that in Dale et al.'s [11], although they have discussed the used of K-Nearest Neighbor, however, they did not specifically report any results of the method. Therefore, we have excluded K-Nearest Neighbor from our comparison as shown in Fig. 7.

The maximum accuracy we got from our experiments was 96.9%, which is higher than the maximum accuracy 91.2% obtained by Dale et al. using similar machine learning algorithms [11] and also the earlier work of Dale et al.'s using Pathologic with 91% accuracy [13]. We believe that, by using an efficient pathway dataset with high number of different representation for each pathway (adding more organisms), and with well-defined set of features could improve the performance of the machine learning classifiers.

V. Conclusion and Future Work

In this research, we have presented a comparative study of our chosen machine learning algorithms with the work that have been established in the metabolic pathway prediction field. Our study was conducted over determining the presence or the absence of a metabolic pathway. We found that only few works addressed the problem of determining the presence or the absence of a metabolic pathway. After that, we started to build a pathway dataset in order to train and evaluate the machine learning algorithms. Our methodology is built upon the four machine learning phases: data acquisition, data preprocessing, training the machine learning algorithms and model evaluation and comparison. Our results shows that the maximum accuracy we got from our experiments was 96.9% given by the Support Vector Machine classifier with and without feature selection methods.

As a future work, there are some points that can be taken into consideration in order to improve this research and these are: (1) Build a pathway dataset with more than 20 organisms in order to increase the number of representation for each pathway. (2) Include all the features that has been defined by Dale et al. or define more features that have better description for each pathway. (3) Develop a deep learning-based prediction for determining the absence or presence of a metabolic pathway.

Contributions

Our work is built upon the work of Prof. Peter D. Karp, the director of Bioinformatics Research Group, SRI International, California, USA, formerly was a Stanford Research Institute to support innovative ideas. Prof. Karp is a pioneer in metabolic pathways studies. He devised a hard-coded algorithm to predict the absence or presence of a pathway using a software tool called Pathologic. He and his colleagues' later studied selected machine learning algorithms (K-Nearest Neighbor, Decision Tree, Logistic Regression and Naive Bayes) to compare with the results of Pathologic. Our work extends Prof. Karp's work by using a new dataset of 20 organisms (as compared to six by them) and also we used an additional algorithm, which is the Support Vector Machine. Our results show an improvements in accuracy as compared to Karp's and his team.

Acknowledgment

Special thanks go to Dr. Ghada Alharbi, vice dean of College of Computer Science and Engineering, for her kind support and guidance. I would also like to express my appreciation to Dr. Liyakath Unisa for her comments and suggestions. I would also like to thank Prof. Peter D. Karp for his valuable advice and for allowing me to base my project on his work.

References

[1] Raval, "Introduction to Biological Networks," Introd. to Biol. Networks, 2018, doi: 10.1201/b14987.

[2] T. Theorell, "Anabolism and catabolism - Antagonistic partners in stress and strain," Scand. J. Work. Environ. Heal. Suppl., no. 6, pp. 136–143, 2008.

[3] B. Junker and S. Falk, Analysis of Biological Networks. A John Wiley & Sons, Inc., Publication, 2008.

[4] L. Liao, S. Kim, and J. Tomb, "Genome Comparisons Based on Profiles of Metabolic Pathways," 2002, pp. 469–476.

[5] R. Overbeek, "WIT: integrated system for high-throughput genome sequence analysis and metabolic reconstruction," Nucleic Acids Res., vol. 28, no. 1, pp. 123–125, 2002, doi: 10.1093/nar/28.1.123.

[6] S. Okuda et al., "KEGG Atlas mapping for global analysis of metabolic pathways.," Nucleic Acids Res., vol. 36, no. Web Server issue, 2008, doi: 10.1093/nar/gkn282.

[7] L. Matthews et al., "Reactome knowledgebase of human biological pathways and processes," Nucleic Acids Res., vol. 37, no. SUPPL. 1, 2009, doi: 10.1093/nar/gkn863.

[8] "Home - Reactome Pathway Database." https://reactome.org/ (accessed Apr. 20, 2019).

[9] G. Kastenmüller, J. Gasteiger, and H. W. Mewes, "An environmental perspective on large-scale genome clustering based on metabolic

capabilities," Bioinformatics, vol. 24, no. 16, pp. 56–62, 2008, doi: 10.1093/bioinformatics/btn302.

[10] G. Kastenmüller, M. E. Schenk, J. Gasteiger, and H. W. Mewes, "Uncovering metabolic pathways relevant to phenotypic traits of microbial genomes," Genome Biol., vol. 10, no. 3, 2009, doi: 10.1186/gb-2009-10-3-r28.

[11] J. M. Dale, L. Popescu, and P. D. Karp, "Machine learning methods for metabolic pathway prediction Supplementary Material Features Used in Machine Learning Predictors," BMC Bioinformatics, pp. 1–13, 2009.

[12] E. Al Daoud, "A new algorithm for Predicting Metabolic Pathways," Int. J. Eng. Sci. Invent., vol. 5, no. 8, pp. 20–24, 2016, [Online]. Available: www.ijesi.org.

[13] S. M. Paley and P. D. Karp, "Evaluation of computational metabolic-pathway predictions for Helicobacter pylori," Bioinformatics, vol. 18, no. 5, pp. 715–724, 2002, doi: 10.1093/bioinformatics/18.5.715.

[14] I. Witten, E. Frank, and M. Hall, Data Mining Practical Machine Learning Tools and Techniques, vol. 54, no. 2. 2011.

[15] P. D. Karp et al., "The BioCyc collection of microbial genomes and metabolic pathways," Brief. Bioinform., no. June, pp. 1–9, 2017, doi: 10.1093/bib/bbx085.

[16] R. Caspi et al., "The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases," Nucleic Acids Res., vol. 44, no. D1, pp. D471–D480, 2016, doi: 10.1093/nar/gkv1164.

[17] I. Guyon, "An Introduction to Variable and Feature Selection," vol. 3, pp. 1157–1182, 2003.

[18] B. Azhagusundari and A. S. Thanamani, "Feature Selection based on Information Gain," Int. J. Innov. TechnolB Azhagusundari Antony Selvadoss Thanamani. 2013. Featur. Sel. based Inf. Gain. Int. J. Innov. Technol. Explor. Eng. (IJITEE), 2(2)18–21.ogy Explor. E, vol. 2, no. 2, pp. 18–21, 2013, doi: 2278-3075.

[19] M. A. Hall, "Correlation-based Feature Selection for Machine Learning," no. April, 1999.

[20] Jecinta Morgan, "Differences Between Supervised Learning and Unsupervised Learning | Difference Between," 2018. http://www.differencebetween.net/technology/differences-between-supervised-learning-and-unsupervised-learning/ (accessed Apr. 18, 2019).

[21] Aditya Mishra, "Metrics to Evaluate your Machine Learning Algorithm," 2018. https://towardsdatascience.com/metrics-to-evaluate-your-machine-learning-algorithm-f10ba6e38234 (accessed Apr. 20, 2019).

[22] H. Jiawei, K. Micheline, and P. Jian, Data Mining, Concepts and Techniques. Elsevier Inc, 2012.

APPENDIX A

ORGANISMS' LIST AND DATABASES

Organism	Database	Version
Escherichia coli K-12 MG1655	EcoCyc	22.6
Arabidopsis thaliana	AraCyc	13.0
Bacteroides thetaiotaomicron VPI-5482	BtheCyc	20.1
Candida albicans SC5314	CalbiCyc	12.0
Chlamydomonas reinhardtii	ChlamyCyc	5.0
Cyanidioschyzon merolae strain 10D	CyanidioCyc	20.0
Emiliania huxleyi CCMP1516	EmilianiaCyc	20.0
Candidatus Evansia muelleri	EvaCyc	1.0.1
Cryptosporidium hominis TU502	HominisCyc	20.5
Homo sapiens	HumanCyc	20.5
Lactobacillus rhamnosus GG	LactorhaCyc	20.5
Candidatus Portiera aleyrodidarum	PabtqvlcCyc	1.0.1
Plasmodium berghei ANKA	PbergheiCyc	20.5
Phaeodactylum tricornutum CCAP 1055/1	PhaeoCyc	20.0
Prevotella copri DSM 18205	PrecopriCyc	20.5
Toxoplasma gondii ME49	ToxoCyc	20.5
Trypanosoma brucei	TrypanoCyc	10.0.1
Saccharomyces cerevisiae S288c	YeastCyc	20.5
Danio rerio	ZfishCyc	18.0
Amycolatopsis mediterranei S699	Amed713604Cyc	19.0

THE TYPES AND DESCRIPTION OF THE 20 FEATURES

Feature	Type	Description
Has-Orphan-Reaction	Boolean	True if the pathway has an orphan reaction.
Has-Spontaneous-Reaction	Boolean	True if the pathway has a spontaneous reaction.
Energy-Pathway	Boolean	True if the pathway is an energy pathway.
Deg-Or-Detox-Pathway	Boolean	True if the pathway is a degradation pathway or a detoxification pathway.
Detoxification-Pathway	Boolean	True if the pathway is a detoxification pathway.
Degradation-Pathway	Boolean	True if the pathway is a degradation pathway.
Biosynthesis-Pathway	Boolean	True if the pathway is a biosynthetic pathway.
Is-Variant	Boolean	True if the pathway is a variant pathway.
Is-Sub-Pathway	Boolean	True if the pathway belongs to any super pathways.
Multiple-Reaction-Pathway	Boolean	True if the pathway has more than one reaction.
Single-Reaction-Pathway	Boolean	True if the pathway has only one reaction.
Num-Reactions	Numeric	Number of reactions in the pathway.
Has-Enzymes	Boolean	True if there are enzymes catalyzing reactions in this pathway.
Num-Enzymes	Numeric	Number of enzymes catalyzing reactions in this pathway.
Enzymes-Per-Reaction	Numeric	Number of enzymes catalyzing reactions in this pathway, divided by number of reactions.
Has-Key-Reactions	Boolean	True if the pathway has key reactions.
Num-Output-Compounds	Numeric	Number of (primary) output compounds of the pathway.
Num-Input-Compounds	Numeric	Number of (primary) input compounds of the pathway.
Num-Input/Output-Compounds	Numeric	Number of (primary) input or output compounds of the pathway.
Num-Initial-Reactions	Numeric	The number of reactions in the pathway that have no predecessors in the pathway.